Jurnal Sarjana Teknik Informatika
Fakultas Teknologi Industri
Universitas Ahmad Dahlan

PENGELOMPOKAN JUDUL PENELITIAN DOSEN MENGGUNAKAN METODE *K-MEANS* DENGAN *COSINE SIMILARITY*

[(1)]Nova Anggraini (1500018196), [(2)] Lisna Zahrotun (0511098401)
Program Studi Teknik Informatika, Fakultas Teknlogi Industri, Universitas Ahmad Dahlan Yogyakarta, Jl. Ringroad Selatan, Banguntapan, Bantul, 55191, Telp : (0274)511830
[(1)]novaanggraini95@gmail.com, [(2)]lisna.zahrotun@tif.uad.ac.id

ABSTRAK

Universitas Ahmad Dahlan (UAD) adalah salah satu Perguruan Tinggi Muhammadiyah yang berada di Provinsi Daerah Istimewa Yogyakarta. UAD memiliki Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) yang menjadi sarana untuk mempublikasikan penelitian dosen. Dosen diwajibkan melakukan kegiatan ilmiah yaitu penelitian dalam memecahkan masalah dengan cara sistematis yang telah ditetapkan untuk mencapai tujuan yang telah dirumuskan. LPPM UAD mengelola judul penelitian dosen dengan cara menyimpan dan mempublikasikan tetapi belum mengelompokkannya. Berdasarkan permasalahan tersebut, maka penelitian ini bermaksud membuat program untuk mengelompokkan judul penelitian dosen berdasarkan kategori penelitiannya.

Metode yang digunakan dalam penelitian ini adalah *K-Means* sedangkan metode pendekatan yang digunakan adalah *Cosine Similarity*. Penelitian ini meliputi tahapan dari *text mining* yaitu *tokenizing, filtering, stemming,* algortima *k-means,* menghitung akurasi menggunakan pengujian *silhouette coefficient*. Data yang digunakan dalam penelitian ini adalah judul penelitian dosen tahun 2015-2017. Penerapan metode *K-Means* digunakan untuk mengelompokkan judul penelitian dosen berdasarkan kategori penelitiannya dengan akurasi klasifikasi yang baik.

Hasil dari penelitian menggunakan 623 data penelitian dosen yang memiliki atribut nama peneliti, prodi, fakultas, judul penelitian dan tahun. Akurasi yang dihasilkan dari penelitian ini menggunakan metode *Silhouette Coefficient* menghasilkan nilai akurasi sebesar 0,6544. Hasil ini tergolong cukup baik karena *range* nilai *silhouette coefficient* dikatakan baik jika nilai semakin mendekati 1 maka semakin baik kualitas kelompoknya.

Kata Kunci : Pengelompokan, *Text Mining, K-Means, Cosine Similarity, Silhouette Coefficient*.

A. PENDAHULUAN

Universitas Ahmad Dahlan (UAD) adalah salah satu Perguruan Tinggi Muhammadiyah yang berada di Provinsi Daerah Istimewa Yogyakarta. Universitas Ahmad Dahlan memiliki Lembaga Penelitian dan Pengabdian Masyarakat (LPPM). LPPM UAD merupakan unsur pelaksana tingkat universitas yang mempunyai tugas mengkoordinasikan, memonitor pelaksanaan kegiatan penelitian, menyelenggarakan kolokium hasil penelitian, dan mengembangkan bidang penelitian yang dilakukan oleh dosen-dosen UAD serta pusat-pusat studi maupun oleh Pusat Pengembangan UAD [1].

Sampai saat ini yang menjadi sarana untuk mempublikasikan penelitian dosen ditangani langsung oleh Lembaga Penelitian dan Pengabdian Masyarakat. Berdasarkan hasil wawancara dengan Bapak Drh. Asep Rustiawan, M.S selaku Sekretaris Lembaga Penelitian dan Pengabdian Masyarakat (LPPM), LPPM telah memiliki sistem yang dapat dilihat oleh pihak dosen untuk mengetahui hasil penelitiannya. Pengarsipan judul penelitian dosen dilakukan pada dua tempat, yaitu di *website* portal.uad.ac.id dan bentuk *file excel.* LPPM tidak mengetahui secara pasti judul-judul penelitian dosen dikarenakan dari pihak LPPM belum mengelompokkan judul-judul penelitian dosen berdasarkan kategori penelitian dan LPPM juga belum mengetahui dosen-dosen yang telah melakukan penelitian dengan judul yang sama. Data-data judul penelitian dosen yang telah ada dapat diidentifikasi kemiripan judul penelitiannya yang dihasilkan dari pengelompokan judul penelitian dosen publikasi LPPM. Dari hasil pengolahan judul dapat dibagi menjadi 4 kategori kelompok yaitu kategori Obat, Makanan, Kesehatan, kategori Pendidikan, kategori Sains dan Teknologi, dan kategori Humaniora sehingga LPPM dapat memberikan informasi topik penelitian kepada dosen yang akan melakukan penelitian berdasarkan judul-judul penelitian sesuai kategori penelitian tahun-tahun sebelumnya. Data penelitian dosen yang diperoleh dari Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) UAD dari tahun 2015-2017 sebanyak 623 judul penelitian. Seperti pada Tabel 1.1. Penelitian Dosen.

Tabel 1.1 : Tabel Penelitian Dosen

No.	Tahun	Jumlah
1.	2015 - 2016	268
2.	2016 - 2017	355
Total		623

Salah satu cara untuk mengetahui kemiripan judul-judul penelitian berdasarkan kategori telah banyak dilakukan pada penelitian sebelumnya. Dilakukan pengelompokan dengan menggunakan metode *K-Means* dengan *Cosine Similarity*. *K-Means* merupakan salah satu metode pengelompokan data yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sedemikian rupa agar data yang berkarakteristik sama dimasukan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Hal ini dilakukan secara bertahap hingga diperoleh kelompok yang tetap [2].

Dengan demikian, melihat dari masalah yang terjadi seperti kantor Lembaga Penelitian dan Pengembangan belum melakukan pengelompokan terhadap judul penelitian dosen berdasarkan kategori penelitiannya, serta kelebihan metode *K-Means* dengan *Cosine Similarity* maka oleh peneliti dilakukan penelitian dengan judul "Pengelompokan Judul Penelitian Dosen Menggunakan Metode *K-Means* dengan *Cosine Similarity*" diharapkan dapat memberikan solusi dari permasalahan pada pengelompokan judul penelitian dosen.

B. KAJIAN PUSTAKA

Penelitian tentang pengelompokan dokumen telah dilakukan oleh Muhammad Sholeh Hudin, M Ali Fauzi, & Sigit Adinugroho (2018) menggunakan metode *k-means clustering* dengan nilai *silhouette* yang dihasilkan 0,483695522 dengan jumlah *cluster k* = 4. Sistem dapat mengelompokkan dokumen dengan menggunakan algoritma *K-Means Clustering* dan *Text Mining*. Sedangkan dalam penelitian Munifah, Syukur, & Supriyanto (2015) membahas mengenai algoritma *K-Means* dengan *Feature Selection Chi Square* yang dilakukan terhadap pengelompokan arsip universitas dengan nilai akurasi sebesar 73,86% pada pembobotan TF-IDF melalui *feature selection chi square*, dengan *time taken* 9 detik.

C. METODE PENELITIAN

1. Text Mining

Text Mining merupakan penerapan konsep dari teknik data *mining* untuk mencari pola dalam teks yang memiliki tujuan untuk mencari informasi yang bermanfaat dengan tujuan tertentu. Proses *text mining* memerlukan beberapa tahap awal untuk mempersiapkan agar teks dapat diubah menjadi lebih terstruktur [3]. Tahapan yang dilakukan pada *preprocessing* terdapat beberapa tahapan *text mining* didalamnya yaitu :

a. Tokenization

Pada *tokenization* terdapat beberapa proses yang harus dilakukan yaitu merubah semua huruf besar menjadi kecil (*text to lowercase*).

b. Stopword

Stopword merupakan proses seleksi terhadap kata-kata yang dihasilkan dari proses *tokenization*, dapat dilakukan dengan algoritma *stoplist* maupun *wordlist*.

c. Stemming

Stemming merupakan proses penghilangan/pemotongan *prefiks* (awalan) dan *sufiks* (akhiran) dari kata dan istilah-istilah dokumen.

2. Pengelompokan

Pengelompokan data menjadi sejumlah kategori juga dapat dilakukan menggunakan metode *clustering*. Berbeda dengan metode klasifikasi, *clustering* mengelompokkan data hanya berdasarkan fitur-fitur yang ada pada data tersebut. Berdasarkan sifat tersebut, *clustering* tidak memerlukan data yang telah diketahui kelasnya. Oleh karena itu, proses pembelajaran pada *clustering* bersifat mandiri, yang sering disebut dengan istilah ***unsupervised learning*** [4]**.**

3. Judul Penelitian

Judul Penelitian adalah suatu kegiatan ilmiah dalam memecahkan masalah dengan cara sistematis yang telah ditetapkan untuk mencapai tujuan yang telah dirumuskan.

Metode penelitian terdiri dari berbagai teknik penelitian apa pun yang kita gunakan, baik kuantitif ataupun kualitatif, haruslah sesuai dengan kerangka teoretis yang kita asumsikan.

4. K-Means

K-Means merupakan metode pengelompokan yang paling popular dan banyak digunakan. Metode ini disusun atas dasar ide yang sederhana. Pada tahap awal ditentukan berapa kelompok yang akan dibentuk. Objek yang diambil secara acak untuk dijadikan titik tengah (*centroid point)* kelompok. Pada tahapan selanjutnya metode *K-Means Clustering* akan melakukan pengulangan langkah-langkah tersebut sampai terjadi tidak ada objek yang dapat dipindahkan atau terjadi kestabilan [6] :

a. Menentukan pendekatan dengan *cosine similarity* untuk mengukur kesamaan antara dua vector dengan mengambil kosinus sudut yang dibuat dua vector.

$$\text{sim}(Xa, Xb) = \cos\theta \frac{Xa.Xb}{\|Xa2\|\|Xb2\|} \quad \textbf{(1)}$$

b. Menentukan jumlah kelompok.

c. Menentukan titik pusat setiap kelompok.

d. Menghitung *centroid* dari data yang ada di masing-masing kelompok serta mengelompokan masing-masing data ke *centroid* tersebut berdasarkan jarak terdekatnya

$$\mathbf{BCV} = \sum_{i,j=0}^{n} \mathbf{d}\ (\mathbf{mi}, \mathbf{mj}) \quad \textbf{(2)}$$

$$\mathbf{WCV} \sum_{i=1}^{n} (\boldsymbol{Jarak\ terdekat\ setiap\ data}) \textbf{(3)}$$

e. Kembali ke step d (Menghitung *centroid* dari data yang ada di masing-masing kelompok) apabila masih ada data yang s.

5. Silhouette Coefficient

Silhouette coefficient digunakan untuk melihat seberapa baik suau objek ditempatkan dalam suatu *cluster* [2]. Tahap perhitungan *silhouette coefficient* adalah sebagai berikut :

a. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu *cluster*

$$\mathbf{a}(\boldsymbol{i}) = \frac{\mathbf{1}}{|\boldsymbol{A}|-\mathbf{1}} \sum \boldsymbol{j} \epsilon \mathbf{A}, \mathbf{j} \neq \mathbf{I}\ \mathbf{d}(\boldsymbol{i}, \boldsymbol{j}) \quad \textbf{(4)}$$

dengan j adalah dokumen lain dalam *cluster* A dan d(i,j) adalah jarak antara dokumen i dengan j

b. Hitung rata-rata dari dokumen i tersebut dengan semua dokumen di *cluster* lain, dan ambil nilai terkecilnya.

$$\mathbf{d}(\boldsymbol{i}, \boldsymbol{C}) = \frac{\mathbf{1}}{|\boldsymbol{A}|-\mathbf{1}} \sum \boldsymbol{j}\ \boldsymbol{\epsilon}\ \mathbf{C}\mathbf{d}(\boldsymbol{i}, \boldsymbol{j}) \quad \textbf{(5)}$$

dimana d(i,C) adalah jarak rata-rata dokumen i dengan semua objek pada *cluster* lain C dimana A ≠ C.

$$\mathbf{b}(\boldsymbol{i}) = \mathbf{min}\ \mathbf{C} \neq \mathbf{A}\ \mathbf{d}(\boldsymbol{i}, \boldsymbol{j}) \quad \textbf{(6)}$$

c. Nilai silhouette coefficient nya adalah :

$$\boldsymbol{s}(\boldsymbol{i}) = \frac{\boldsymbol{b}(\boldsymbol{i}) - \mathbf{a}(\boldsymbol{i})}{\mathbf{ma\,x}\{\mathbf{a}(\boldsymbol{i}), \boldsymbol{b}(\boldsymbol{i})\}}$$

(7)

D. HASIL DAN PEMBAHASAN

1. Load Data

Data Judul Penelitian yang sudah dikumpulkan lalu diberi ID per dokumen, setiap judul dianggap sebagai satu dokumen seperti pada Tabel 2 : Load Data

Tabel 2 : Load Data

ID	JUDUL PENELITIAN
D0	Hubungan Kekerabatan Fenetik Suku Asteraceae di Yogyakarta
D1	Hubungan Perilaku Keimanan, Ihsan, Sabar,dan Syukur dengan Kebahagiaan dan Kebermaknaan Hidup
D2	Intervensi Kekerasan Terhadap Anak Usia Dini di Yogyakarta
D3	Edukasi Dampak dan Bahaya Rokok pada Siswa Sekolah Dasar di Dusun Bodon, Banguntapan
D4	Uji Coba Modul Konseling Farmasis (KSF) untuk Pasien Diabetes Mellitus (DM) Hipertensi
D5	HUBUNGAN KEPATUHAN TERAPI OBAT TERHADAP KUALITAS HIDUP PADA PASIEN HIPERTENSI DI PUSKESMAS MERGANGSAN YOGYAKARTA
D6	Analisis Biaya Medis Langsung Pengobatan Stroke di Rumah Sakit PKU Muhammadiyah Bantul Yogyakarta
D7	Sistem Navigasi Robot dalam Ruangan
D8	PERANCANGAN VISUALISASI INFORMASI UNTUK SISTEM EVALUASI GURU
D9	Perancangan Antarmuka Layanan Perpustakaan Berbasis RFID di PSB UAD

2. Prepocessing

Proses setelah *load* data adalah proses *prepocessing.* Didalam proses *preprocessing* terdapat 3 tahapan yaitu, *tokenizing,* untuk pemotongan *string input* berdasarkan tiap kata yang menyusunnya, *filtering* untuk membuang kata-kata yang dianggap tidak penting seperti kata hubung dan lain-lain, dan *stemming* untuk menghilangkan atau memotong *prefix* (awalan) dan *suffixs* (akhiran) dari kata dan istilah-istilah dokumen menjadi kata dasar. Hasil proses *prepocessing* dapat dilihat pada Tabel 3 : Hasil Proses *Prepocessing* Judul Penelitian Dosen.

Tabel 3 : Hasil Proses *Prepocessing* Judul Penelitian Dosen

ID	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
D0	hubung	kerabat	fenetik	suku	asteraceae	yogyakarta						
D1	hubung	perilaku	iman	ihsan	sabar	syukur	bahagia	makna	hidup			
D2	intervensi	keras	anak	usia	dini	yogyakarta						
D3	edukasi	dampak	bahaya	rokok	siswa	sekolah	dasar	dusun	bodon	banguntapan		
D4	uji	coba	modul	konseling	farmasis	ksf	pasien	diabetes	mellitus	dm	hipertensi	
D5	hubung	patuh	terapi	obat	kualitas	hidup	pasien	hipertensi	puskesmas	mergangsan	yogyakarta	
D6	analisis	biaya	medis	langsung	obat	stroke	rumah	sakit	pku	muhammadiyah	bantul	yogyakarta
D7	sistem	navigasi	robot	ruang								
D8	ancang	visualisasi	informasi	sistem	evaluasi	guru						
D9	ancang	antarmuka	layan	pustaka	basis	rfid	psb	uad				

3. K-Means

Pada penelitian ini contoh data yang diambil yaitu 10 data Penelitian Dosen UAD. Tahapan proses *K-Means* antara lain sebagai berikut :

a. Proses awal mengukur kesamaan antara dokumen satu dengan yang lain menggunakan *cosine similarity* rumus *cosine* dapat dilihat pada persamaan (1).

$$\text{Cos (D0,D0)} = \frac{(1.1)+(1.1)+(1.1)+(1.1)+(1.1)+(1.1)}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{6}{\sqrt{6}\times\sqrt{6}} = \frac{6}{2.449\times 2.449} = \frac{6}{5.997} = 1$$

$$\text{Cos (D0,D1)} = \frac{(1.1)+0+0+0+0+0+0+0+0+0+0+0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{1}{\sqrt{6}\times\sqrt{9}} = \frac{1}{2.449\times 3} = \frac{1}{7.374} = 0.136$$

Hasil proses *cosine similarity* terdapat pada Tabel 4 : Hasil Proses *Cosine Similarity*

Tabel 4 : Hasil Proses *Cosine Similarity*

ID	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
D0	1	0.136	0.167	0	0	0.246	0.118	0	0	0
D1	0.136	1	0	0	0	0.201	0	0	0	0
D2	0.167	0	1	0	0	0.123	0.118	0	0	0
D3	0	0	0	1	0	0	0	0	0	0
D4	0	0	0	0	1	0.182	0	0	0	0
D5	0.246	0.201	0.123	0	0.182	1	0.174	0	0	0
D6	0.118	0	0.118	0	0	0.174	1	0	0	0
D7	0	0	0	0	0	0	0	1	0.204	0
D8	0	0	0	0	0	0	0	0.204	1	0.144
D9	0	0	0	0	0	0	0	0	0.144	1

b. Setelah mendapatkan hasil *cosine similarity* kemudian dilanjutkan dengan menentukan jumlah kelompok sebanyak 4 *cluster* ditentukan secara acak.

c. Diketahui titik pusatnya 4 yaitu D0, D2, D5, dan D8. Dalam menentukan jarak terdekat dengan membandingkan nilai terdekat dari C1, C2, C3, dan C4 pada setiap 10 dokumen. Rumus dapat dilihat pada persamaan (1).

Cos (D0,C1) = $\frac{(1.1)+(1.1)+(1.1)+(1.1)+(1.1)+(1.1)}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2\times}\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{6}{\sqrt{6}\times\sqrt{6}} = \frac{6}{2.449\times 2.449} = \frac{6}{5.997} = 1$

Cos (D0,C2) = $\frac{(1.1)+0+0+0+0+0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2\times}\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{1}{\sqrt{6}\times\sqrt{6}} = \frac{1}{2.449\times 2.449} = \frac{1}{5.997} = 0.167$

Maka diketahui *cluster* 1 (C1) terdapat 1 dokumen yaitu: D0, *cluster* 2 (C2) terdapat 1 dokumen yaitu: D2, *cluster* 3 (C3) terdapat 4 dokumen yaitu: D1, D4, D5, dan D6, *cluster* 4 (C4) terdapat 3 dokumen yaitu: D7, D8, dan D9. Hasil *cluster* awal setelah menentukan titik pusat secara random terdapat pada Tabel 5 : Hasil *Cluster* Awal

Tabel 5 : Hasil *Cluster* Awal

Dn	C1	C2	C3	C4	Hasil Similarity
D0	1	0.167	0.246	0	C1
D1	0.136	0	0.201	0	C3
D2	0.167	1	0.123	0	C2
D3	0	0	0	0	-
D4	0	0	0.181	0	C3
D5	0.246	0.123	1	0	C3
D6	0.117	0.117	0.174	0	C3
D7	0	0	0	0.204	C4
D8	0	0	0	1	C4
D9	0	0	0	0.144	C4

d. Pada langkah ini dihitung pula rasio besaran *Between Cluster Variation* (BCV) dengan *Within Cluster Variation* (WCV), rumus BCV dapat dilihat pada persamaan (2) dan rumus WCV dapat dilihat pada persamaan (3) seperti berikut:

BCV = d(C1,C2) + d(C1,C3) + d(C1,C4) + d(C2,C3) + d(C2,C4) + d(C3,C4)

1. (C1,C2)

C1 = hubung kerabat fenetik suku asteraceae yogyakarta

C2 = intervensi keras anak usia dini yogyakarta

(C1,C2) = $\frac{(1.1)+0+0+0+0+0+0+0+0+0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2}\,x\,\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{1}{\sqrt{6}\,x\sqrt{6}} = \frac{1}{2.449\,x\,2.449} = \frac{1}{5.997} = 0.167$

2. (C1,C3)

C1 = hubung kerabat fenetik suku asteraceae yogyakarta

C3 = hubung patuh terapi obat kualitas hidup pasien hipertensi puskesmas mergangsan yogyakarta

(C1,C3) = $\frac{(1.1)+(1.1)+0+0+0+0+0+0+0+0+0+0+0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2}\,x\,\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2}} = \frac{2}{\sqrt{6}\,x\sqrt{11}} = \frac{2}{2.449\,x\,3.316} = \frac{2}{8.120} = 0.246$

BCV = 0.167 + 0.246 + 0 + 0.123 + 0 + 0 = 0.536

Setelah mendapatkan hasil BCV kemudian menghitung WCV

WCV = $1^2 + 0.201^2 + 1^2 + 0^2 + 0.181^2 + 1^2 + 0.174^2 + 0.204^2 + 1^2 + 0.144^2$

= 1 + 0.040 + 1 + 0 + 0.032 + 1 + 0.030 + 0.041 + 1 + 0.020

= 4.163

Sehingga besaran rasio adalah :

$\frac{BCV}{WCV} = \frac{0.536}{4.163} = 0.128$

e. Untuk menghitung nilai Iterasi dibutuhkan titik pusat baru yang telah dibentuk dari anggota kelompok *cluster* awal. Hasil Iterasi dapat dilihat pada Tabel 6 : Hasil *Cluster* Iterasi

Tabel 6 : Hasil *Cluster* Iterasi

Dn	C1	C2	C3	C4	Hasil Similarity
D0	1	0.167	0.333	0	C1
D1	0.136	0	0.272	0	C3
D2	0.167	1	0.167	0	C2
D3	0	0	0	0	-
D4	0	0	0.246	0	C3
D5	0.246	0.123	0.738	0	C3
D6	0.117	0.117	0.235	0	C3
D7	0	0	0	0. 204	C4
D8	0	0	0	0.333	C4
D9	0	0	0	0.144	C4

Dari Tabel 6 didapatkan keangotaan sebagai berikut :

a. Kelompok 1 (atau C1) = {D1}
b. Kelompok 2 (atau C2) = {D2}
c. Kelompok 3 (atau C3) = {D1, D4, D5, D6}
d. Kelompok 4 (atau C4) = {D7, D8, D9}

f. Sehingga terbentuk sebuah *cluster* dengan C=4 dari proses pengelompokan menggunakan metode *k-means* yang dapat dilihat pada Tabel 7 : Hasil *Cluster*

Tabel 7 : Hasil *Cluster*

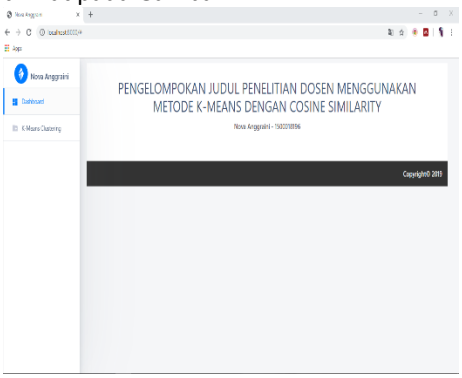
Cluster	Judul Penelitian
Cluster 1	Hubung kerabat fenetik suku asteraceae Yogyakarta
Cluster 2	Intervensi keras anak usia dini Yogyakarta
Cluster 3	Hubung perilaku iman ihsan sabar syukur bahagia makna hidup
	Uji coba modul konseling farmasis ksf pasien diabetes mellitus dm hipertensi
	Hubung patuh terapi obat kualitas hidup pasien hipertensi puskesmas mergangsan Yogyakarta
	Analisis biaya medis langsung obat stroke rumah sakit pku muhammadiyah bantul Yogyakarta
Cluster 4	Sistem navigasi robot ruang
	Ancang visualisasi informasi sistem evaluasi guru
	Ancang antarmuka layan pustaka basis rfid psb uad

4. Implementasi

Implementasi dari Program ini menggunakan bahasa pemrograman Python yang digabungkan dengan Framework Flask. Menghasilkan aplikasi *text mining* berbasis web dengan implementasi sebagai berikut:

a. *Dashboard*

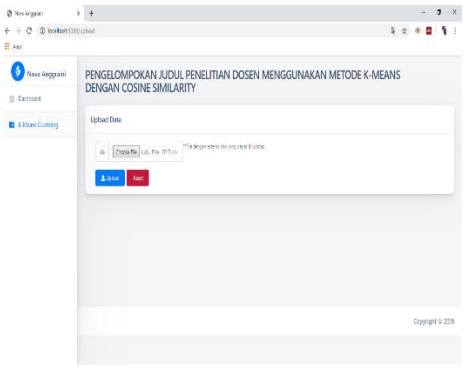
Implementasi untuk *dashboard* terdiri dari menu *Dashboard* dan *K-Means Clustering.* Pada menu *dashboard* terdapat tampilan nama judul pada sistem. Tampilan dapat dilihat pada Gambar 1.



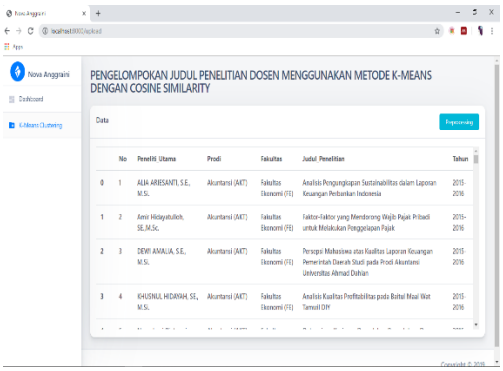
Gambar 1 : *Interface Dashboard*

b. *Load* Data dan Tampil Data

Implementasi untuk *load* data dan tampil data terdapat sistem *load* data untuk mengupload data yang berekstensi .xlsx dan setelah melakukan proses *upload* maka akan menampilkan data yang telah di *upload* kedalam sistem. Tampilan dapat dilihat pada Gambar 2. Dan Gambar 3.



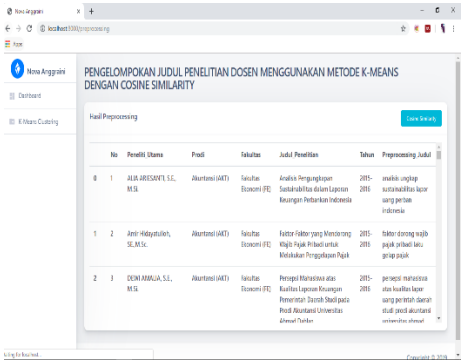
Gambar 2 : *Interface load* data



Gambar 3 : *Interface* Tampil Data

c. *Prepocessing* Data

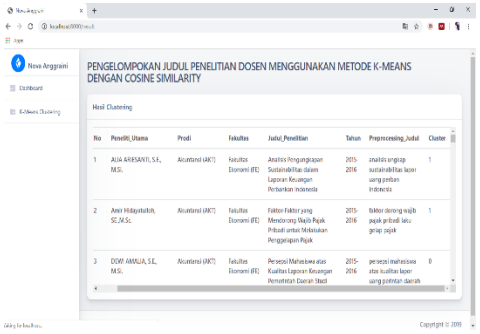
Implementasi *preprocessing* data menampilkan data yang telah di *upload* dan dilakukan proses *preprocessing* data meliputi (*tokenizing, filtering,* dan *stemming*) sehingga menambah kolom baru yaitu *Preprocessing*_Judul. Tampilan dapat dilihat pada Gambar 4.



Gambar 4 : *Interface Prepocessing* Data

d. *K-Means*

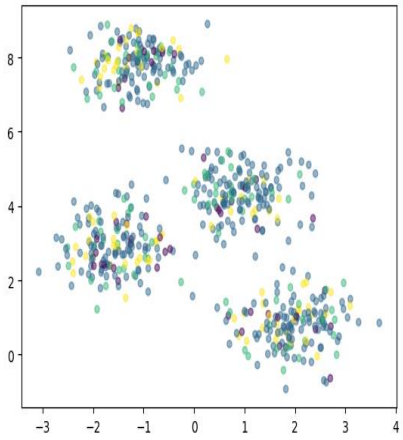
Implementasi untuk *K-Means* pada sistem menampilkan hasil proses yang telah dilalui sebelumnya dan menambah kolom *Cluster* pada sistem. Tampilan dapat dilihat pada Gambar 5.



Gambar 5 : *Interface K-Means*

e. Plot *Clustering*

Implementasi untuk Plot *Clustering* sistem terdapat 4 *blobs* yang berwarna ungu, biru, tosca, dan kuning. Hasil Plot dapat dilihat pada Gambar 6.



Gambar 6 : Interface Grafik

5. Pengujian Akurasi

Pengujian akurasi menggunakan *silhouette coefficient* dilakukan untuk mengetahui baik buruknya kelompok yang dihasilkan dari proses pengelompokan dengan *K-Means*. Jika hasil *silhouette coefficient* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Sedangkan, jika hasil *silhouette coefficient* semakin jauh dari 1, maka semakin buruk kualitas kelompoknya. Hasil yang didapatkan dari pengujian dengan menggunakan *silhouette coefficient* dari 10 data, didapatkan hasil sebagai berikut:

a. Hitung rata-rata jarak dari suatu dokumen misalkan *i* dengan semua dokumen lain yang berada dalam satu *cluster.* Rumus menghitung nilai a(i) dapat dilihat pada persamaan (4). Maka didapatkan hasil yang dapat dilihat pada Tabel 8 :

Tabel 8 : Hasil Perhitungan Nilai a(i)

a(i)	Hasil
a(0)	0
a(1)	0.406
a(2)	0
a(3)	0
a(4)	0.415
a(5)	0.245
a(6)	0.418
a(7)	0.223
a(8)	0.174
a(9)	0.268

b. Hitung rata-rata jarak dari dokumen *i* tersebut dengan semua dokumen di *cluster* lain, dan ambil nilai terkecilnya. Nilai d(I,C) yang dihasilkan akan memiliki 3 nilai karena jumlah *cluster* pada penelitian ini memiliki 4 *cluster*. Setelah itu akan diambil nilai minimum dari 3 nilai d(i,C) yang dihasilkan untuk mendapatkan nilai b(i). Rumus menghitung nilai d(I,C) dapat dilihat pada persamaan (5) dan nilai b(i) dapat dilihat pada persamaan (6).

Tabel 9 : Hasil Perhitungan Nilai b(i)

d(i,C)	Hasil	d(i,C)	Hasil	d(i,C)	Hasil	b(i)	Hasil
d(0,1)	0	d(0,2)	0	d(0,3)	0	d(0)	0
d(1,1)	0.499	d(1,2)	0.444	d(1,3)	0.227	d(1)	0.227
d(2,1)	0	d(2,2)	0	d(2,3)	0	d(2)	0
d(3,1)	0	d(3,2)	0	d(3,3)	0	d(3)	0
d(4,1)	0.499	d(4,2)	0.444	d(4,3)	0.227	d(4)	0.227
d(5,1)	0.499	d(5,2)	0.444	d(5,3)	0.227	d(5)	0.227
d(6,1)	0.499	d(6,2)	0.444	d(6,3)	0.227	d(6)	0.227
d(7,1)	0.749	d(7,2)	0.666	d(7,3)	1.115	d(7)	0.666
d(8,1)	0.749	d(8,2)	0.666	d(8,3)	1.115	d(8)	0.666
d(9,1)	0.749	d(9,2)	0.666	d(9,3)	1.115	d(9)	0.666

c. Hitung nilai *silhouette coefficient* s(i) dengan mencari nilai maksimal dari dari a(i) dan b(i). Rumus perhitungan s(i) dapat dilihat pada persamaan (7).

Tabel 10 : Hasil Perhitungan Nilai s(i)

s(i)	Hasil
s(0)	0
s(1)	-0.440
s(2)	0
s(3)	0
s(4)	-0.453
s(5)	-0.073
s(6)	-0.456
s(7)	0.657
s(8)	0.738
s(9)	0.597

Setelah didapat hasil dari s(i) masing-masing data, maka nilai rata-rata yang digunakan sebagai nilai *slihoutte coefficient* dari hasil *K-Means clustering* dengan *cosine similarity* adalah 0.57 hasil tersebut tergolong tinggi, karena *silhouette coefficient* memiliki *range* nilai antara 0 - 1. Semakin mendekati 1 hasil kelompoknya semakin baik.

E. KESIMPULAN

Berdasarkan hasil penelitian "Pengelompokan Judul Penelitian Dosen Menggunakan Metode *K-Means* dengan *Cosine Similarity*" dapat ditarik kesimpulan sebagai berikut :

1. Telah dibuat aplikasi dengan bahasa pemrograman Python yang mampu mengelompokan judul penelitian.
2. Uji *silhouette coefficient* yang dilakukan pada program "Pengelompokan Judul Penelitian Dosen Menggunakan Metode *K-Means* dengan *Cosine Similarity*" menggunakan 10 data menunjukan hasil sebesar 0.57 dengan *range cluster* = 4. Dan pola kelompok yang dihasilkan dari 623 data dan saat dibagi menjadi 4 kelompok menghasilkan nilai *silhouette coefficient* sebesar 0.6544. Hasil tersebut tergolong cukup baik. 4 *cluster* yang dihasilkan meliputi :
   C1 = Obat, Makanan, & Kesehatan
   C2 = Pendidikan,
   C3 = Sains dan Teknologi
   C4 = Humaniora
3. Pola yang dihasilkan dari 623 data dengan *range cluster* = 4 menunjukan *blobs* plot pada *cluster* 1 kategori Obat, Makanan, dan Kesehatan sebanyak 88 data, *cluster* 2 kategori Pendidikan sebanyak 105 data, *cluster* 3 kategori Sains dan Teknologi sebanyak 382 data dan *cluster* 4 kategori Humaniora sebanyak 48 data. Dari plots yang dihasilkan kategori *cluster* yang dominan atau yang paling banyak adalah *cluster* 3 Sains dan Teknologi.

F. DAFTAR PUSTAKA

[1] Widodo Rustiawan, A Dkk, S (2016). 'Rencana Induk Penelitian (RIP) Universitas Ahmad Dahlan Tahun 2016-2021', pp. 4-5. Available at: http://lpp.uad.ac.id/wp-content/uploads/2016/12/RIP-UAD-2016-2017-KOMPLIT.pdf (Accessed: 4 September 2018)

[2] Muhammad Sholeh Hudin, M. A., Fauzi, S dan Adinugroho (2018) 'Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)', *Pengembangan Teknologi Informasi dan Ilmu*.

[3] U. N. Surabaya, "Universitas Negeri Surabaya."

[4] Adinugroho, S. dan Sari, Y. A. (2018) Implementasi Data Mining Menggunakan *WEKA*. Pertama. Malang L UB Press.

[5] Muh. Fitrah, M. P. dan Dr. Luthfiyah, M.A. (2017) Metodelogi Penelitian; Penelitian Kualitatif, Tindak Kelas Studi Kasus. Pertama. Sukabumi: CV Jejak.

[6] Prilianti, K.R dan Wijaya, H. (2014) 'Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering', Jurnal Cybermatika, 2(1), oo. 1-6.